Request for Information:
GFS/DFS File Systems

October 30, 2000

1.0 Background

The following is intended to serve as guidance for responders to this GFS/DFS PathForward Request for Information. The GFS/DFS PathForward wishes to accelerate global file system development activities. It is the desire of this project to find promising global file system product development projects that could benefit from additional development funded by ASCI. An ASCI file system may be characterized as secure, extremely scalable and able to support complex multiple supercomputer sites. It is likely that most global file system development projects have not considered all the ramifications of such an environment, and it is hoped that one or more of these projects would desire to work with the ASCI project to add scalability and security capability to help the product in question to scale to enormous proportions in a secure and manageable way. The intent of this PathForward process is to bring selected critical technologies or technology enhancements to the marketplace more quickly than they might normally appear. For an example of the PathForward process, please see the computer interconnect technologies PathForward webpage (http://www.llnl.gov/asci-pathforward/).

We recognize that this RFI touches upon areas that may be departmentalized into different teams in some large organizations, and that some small organizations may be unable to address all of our needs. Therefore, responses are encouraged which address ***any*** (as opposed to ***all***) of the five desired characteristics outlined in 2.1 through 2.5. We would prefer a single end-to-end solution for all of our needs, but we are prepared to act as system integrator in the scenario that no firm is able to address all of our needs in the given timeframe. Responses should address the proposed technology development path. Responses which indicate development timeframes of 1 to 3 years will be considered, however responses capable of delivering a series of increasingly capable systems every 12-18 months are preferred.

Before a file system PathForward can be approved, we need to determine if there is sufficient interest from firms, and that a file system PathForward process would provide benefit to the national labs. This first step, soliciting information from interested firms, will help us take the pulse of the industry with respect to our desired technologies *with minimal effort from potential participants*.

2.0 Desired Technologies

There are a number of file system technologies which need to be provided and possibly significantly improved if the national labs are to carry out their defined mission for the next few years. Moreover, we are unaware of new products and/or features intended to address all of these critical file system technologies. We have divided these critical desired technologies into five areas: Global File System; Scalable Access; Scalable Management Facilities; Security; and WAN Access.

Throughout this set of desired technologies, as we focus on the global aspects, unique security, and special scalability requirements of ASCI-scale systems, the usual requirements of any file system remain in place. For example, requirements such as posix compliance, standard locking mechanisms, persistence, integrity, and stability will be assumed.

2.1 Global File System

Due to the nature of the ASCI tri-lab mission, we have need for file systems which are heterogeneous and global.

There are several aspects associated with being a Global File System in our view:

- The name space of the file system must be global. Global means that it is possible to construct a view of the distributed file system hierarchical name space that is identical simultaneously at multiple participating sites and clients. In other words, subject to a site's administrative constraints, it should be possible to provide seamless name space translation to another, participating, site's name space. Further, at least one fully qualified path name to any file or directory object should be identical from any client anywhere without requiring the user to know the actual location of the data or metadata..
- Access to the file system must be heterogeneous. The ASCI environment is made up of Linux, Tru64, AIX, IRIX, and Windows/NT/2000 operating systems. Within each of these operating system environments, many levels of the operating system including older versions and extremely new (including beta) releases of operating systems co-exist. Unlike in many classical business environments, where lagging operating system version support for software is not a penalty due to the need for extreme stability, in the ASCI environment, frequently very young OS levels are required for scalability features and other extreme environmental support features. Given this large set of OS support required, a palatable story for how OS support for the global file system client will be provided (e.g., consortia of OS vendors, sample client code with public source, full open client protocols, etc.).
- It is also desirable that any global file system be exportable via NFS or CIFS to gain an even wider usage base all the way to tens of thousands of common desk top technologies in all flavors of Unix, Windows, and MacIntosh.

2.2 Scalable Access

It is expected that many of the global file system projects considered in the PathForward activity will not have the extreme scalability in both data and meta-data operations in mind. This area is an example of an area where PathForward funding for enhancements and other funded development activities could benefit firms in the file system market.

Scalability is a primary file system concern and there are a few important desired aspects for scalability:

- Capacity: Today, each ASCI site has local file systems which reach sizes of 75 TB in support of around 3,000 users. A single GFS would unify these file systems. Future file systems would scale up with upcoming machines.

- Metadata Performance: Metadata operations such as file creates and so forth need to scale with the number of processors present in ASCI platforms. That is, metadata operations should not be serialized.

- Aggregate Read/Write Bandwidth: Aggregate read/write bandwidth should scale with processor speed and memory size. The "Scalable Access" goal is to have one name space that can be accessed by both compute engines and visualization engines at 100's of MB/sec in the near term, with increasing bandwidths in out years.

As the disparity between processor speeds and I/O interfaces increases, the need to scale activities via parallel access to multiple devices becomes more critical.

- Data Movement Scalability: Data must be able to move between multiple media sources and sinks in parallel. Transfers between multiple clients and multiple independent file objects must be able to proceed in a fashion that minimizes mutual interference [that is high throughput for independent, concurrent file accesses]. As well, transfers to a single file from multiple processes should have minimum interference in a similar fashion [maximum throughput with concurrent access from multiple processes to the same file]. Ideally, benchmarks of the aggregate throughput to multiple files by independent processes should demonstrate linear scalability up to the limit imposed by the underlying system software and hardware. Similarly, coordinated access to a single file by multiple processes should be able to demonstrate linear scalability when access is made to non-overlapping allocation units. Furthermore, support for parallel transfers must support environments where there are one or more file system clients per "SMP compute platform" or on cooperating multiple SMPs. It is vital that the underlying design of the global file system project not preclude scaling given the appropriate hardware infrastructure, and desirable that the global file system enhance the ease in which scalability can occur.

- Meta-Data Scalability: Just as having the ability to scale data movement, scaling meta-data operations is also an important desired technology. A file system may be asked to insert ten thousand files with one request (or set of requests in a very short time). It is important that scaling to handle large numbers of meta-data transactions such as file insert, file delete, etc., be possible. It is vital that the underlying design of the global file system not preclude meta-data scaling, and it is highly desirable that the global file system enhance the ease in which meta-data scalability can occur.

- It is desirable that the design of the global file system promote or not preclude the idea of minimizing the use of heavy weight stack processing on the clients of the file system. Emerging System or Storage Area Networking (SAN) technology could be used in this light. It is desirable that standards driven protocols and access api's be used if this feature is to be utilized.

2.3 Scalable Management Facilities

It is expected that many of the global file system projects considered in the PathForward activity will not have the needed management features mind. This area is an example of an area where PathForward funding for enhancements and other funded development activities could benefit firms in the file system market.

Historically, disk drives at the national labs have always been "hosted" by a computer with special hardware (I/O adapters) to drive the disks. We want the data on each disk to be available to approved remote nodes too, but when data is read or written from a remote node it is very undesirable to require the data to pass through a heavy software stack on the "host" to the disk. Emerging technologies such as System Area Networks (SAN) and Network Attached Storage (NAS) remove this problem, but introduce another: software is needed to specify which nodes can access which blocks on the disks. Furthermore, the type of access (read, write, delete, create, ...) needs to be managed.

It is desirable that future file systems must be able to exploit, in a flexible and extensible manner, the SANs that will be an integral part of the ASCI sites. This integration could include both network integration and interface to network integration, for instance, employing things like ST or other OS bypass mechanisms. However, the choices must not be limiting in nature. The server software and client file system implementations must be able to make use of transports not yet developed and, potentially, only available at a particular site. This may be done through a middle-ware communications layer, pluggable modules or standard, transport independent interfaces, for instance. (I'm not sure about this one but it doesn't seem to read quite right. It seems we have one more "standard" than we might need.)

It is also desirable that the management of a very scalable global file system be scalable as well. In other words, it is important that management overhead of a global file system not increase linearly as the size of the file system grows; this includes meta-data growth, data movement bandwidth growth, and total storage capacity growth.

2.4 Security

It is expected that many of the global file system projects considered in the PathForward activity will not have the needed security features mind. This area is an example of an area where PathForward funding for enhancements and other funded development activities could benefit firms in the file system market.

We need adequate measures to enforce our need-to-know orders, and adequate logging to assess external and internal attempts to thwart such measures.

We require fine-grain access control mechanisms and auditing mechanisms to support a need-to-know sharing and protection model, authorization mechanisms that will integrated with our current authentication and security infrastructure, and data protection and integrity for information in transit.

Existing security-related technologies (e.g., Access Control Lists, shared secret key systems ala *Kerberos*, public key systems ala *Entrust*, transport-based data protection and security ala *Ipsec*) provide some aspects of a fine-grain need-to-know file system. However, a number of issues still remain before a GFS file system can be trusted to provide fine grain inter-site security with a possible “insider threat”.

2.5 WAN Access

It is expected that many of the global file system projects considered in the PathForward activity will not have the needed WAN access features mind. This area is an example of an area where PathForward funding for enhancements and other funded development activities could benefit firms in the file system market.

We must connect remote ASCI sites into one large collective inter-site. Each local site has different administrative policies, different people who have privileges to make local modifications, and possibly different local names for the same user or object. Furthermore, using the "inter-site" should be intuitive. At one time, the DFS product from TransArc was believed to be a possible solution for many of the file system issues. Unfortunately, DFS support is waning.

Resources at each ASCI site, in addition to operating and being managed as autonomous units, must inter-operate between sites and support remote access within several contexts. These include the need for uniform naming, seamless access with minimal differentiation between local and remote resources, authorization controls to uniquely and properly grant privilege to locally and remotely authenticated entities, wide platform availability and the need for strong and sustained industry support.

Using tools like ftp, users can access data on remote resources but this generally results in the creation of a local copy. This becomes a maintenance nightmare. Moreover, it is extremely difficult to maintain the proper authorization controls for multiple copies of the same data or even a single copy of the data if the local site's authorization controls are not able to inter-operate with a remote site's security infrastructure.

Approaches to beef up SANs to work in a WAN environment, or to influence the design of NFS v4 appear like interesting PathForward possibilities.

Appendix A – RFI Contact List

Role	Person	Email	Phones	Mail
Primary RFI Contact	Ann Huber	huber2@llnl.gov	Voice: 925-422-6564 Fax: 925-423-8019	Mail Station L-550 LLNL P.O. Box 808 Livermore, CA 94551-0808
Secondary RFI Contact	Kelly Miller	miller66@llnl.gov	Voice: 925-422-9062 Fax: 925-423-8019	Mail Station L-550 LLNL P.O. Box 808 Livermore, CA 94551-0808
Technical Advisory Committee	Gary Grider	ggrider@lanl.gov	Voice: 505-665-9077 Fax: 505-665-6333	Mailstop B272 CIC-8 LANL Los Alamos, NM 87545
Technical Advisory Committee	Terry Jones	trj@llnl.gov	Voice: 925-423-9834 Fax: 925-423-8704	Mail Station L-561 LLNL P.O. Box 808 Livermore, CA 94551-0808
Technical Advisory Committee	Lee Ward	lward@sandia.gov	Voice: 505-844-9545	Mailstop 1110 SNL P.O. Box 5800 Albuquerque, NM 87185-1110

Glossary

API	Application Programming Interface – The functions and prototypes that a given software layer can program to.
ASCI	Accelerated Strategic Computing Initiative – A U.S. Government funded program which aims to make predictive simulation possible; stimulate the U.S. computer manufacturing industry to create more powerful, high-end supercomputing capability required by these applications; create a computational infrastructure and operating environment that makes these capabilities accessible and usable.
Blue-Mountain	The initial three machines purchased under the ASCI program were Red (a large Intel Teraflops at Sandia National Labs), Blue-Mountain (a large cluster of SGI Origin systems at Los Alamos National Lab), and Blue-Pacific (a large IBM SP2 at Lawrence Livermore National Lab).
Blue-Pacific	The initial three machines purchased under the ASCI program were Red (a large Intel Teraflops at Sandia National Labs), Blue-Mountain (a large cluster of SGI Origin systems at Los Alamos National Lab), and Blue-Pacific (a large IBM SP2 at Lawrence Livermore National Lab).
DFS	Distributed File System: A file system which may be mounted by multiple clients distributed over a computer network. An example of a DFS is NFS. Note: In this document, we use DFS generically for any distributed file system in this document – any reference to the TransArc product with the same name will be clearly specified.
DisCom^2	Distance and Distributed Computing and Communication: DisCom^2 is an ASCI project intended to deliver key computing and communications technologies to efficiently integrate distributed resources with high-end computing resources at a distance.
DMF	DMF is a project initiated at the three labs to address shareable files between sites. It is intended to be a comprehensive solution for fast, portable, serial and parallel I/O providing data share-ability and application & tool interoperability for scientific data.
DOE	The United States Department of Energy (http://www.doe.gov).
DP-10	LANL, LLNL, and Sandia are under the umbrella of the Department of Energy. The work proposed in this paper would be funded by the Defense Program (DP) of the DOE. The Defense Program is divided up into several subprograms (e.g., DP-10, DP-20, DP-30, DP-40, and DP-50). This work falls into DP-10 which is "Strategic Computing and Simulation."
GFS	Global File System: (Not to be confused with Univ. of Minnesota's GFS): A file system that provides a single unified name space across multiple (possibly heterogeneous) platforms.
HDF5	A low level I/O API and file format. Provides a full-featured I/O system enabling data subsetting, portability, etc.
NAP	Network Attached Peripheral: Individual NAS components.
NAS	Network Attached Storage: Devices that provide storage services on an internet or intranet.
NASD	Network Attached Secure Disks: A NAP with added security features.
PathForward	Funding delivered to industry to accelerate possible commercial solutions for ASCI needs. There are software PathForwards and hardware PathForwards.

POSIX	Portable Operating Systems Interface – An international standard developed by the IEEE and adopted by the ISO. Provides UNIX users with an international harmonized standard for operating system interfaces.
PSE	Problem Solving Environment
RAID	Redundant Array of Independent Disks – striping of a stream of data onto multiple disks usually with some kind of hardware generated parity stripe.
RAIT	Redundant Array of Independent Tapes – striping of a stream of data onto multiple tape drives usually with some kind of hardware generated parity stripe.
Red	The initial three machines purchased under the ASCI program were Red (a large Intel Teraflops at Sandia National Labs), Blue-Mountain (a large cluster of SGI Origin systems at Los Alamos National Lab), and Blue-Pacific (a large IBM SP2 at Lawrence Livermore National Lab).
RFI	Request For Information: A call for information. In contrast to an RFP, an RFI does not require a detailed proposal. That is, it does not generally require a thorough itemized project plan, a thorough itemized funding plan, nor any documentation on anticipated contractual terms and conditions.
RFP	Request For Proposal: A formal request to potential suppliers for a proposed solution, including a technical scope and pricing. RFPs are used for both purchasing a sophisticated item and for funding directed R&D. See also RFI.
SAN	Storage Area Networks. A dedicated network wherein general host(s) access either NAS or NAPs.
SCCD	Scientific Computing and Communications Department – The supercomputer center at LLNL.
SDM	Scientific Data Management. SDM is a subproject with the VIEWs project to develop an environment that allows scientists to store, retrieve, search and reduce data within the natural context of their work. This framework integrates scientific data models, commercial databases, mass storage systems, networking and computing infrastructure, and intelligent post-processing to provide assistance in managing the complexity and scale of ASCI data.
Tri-lab	Refers to the three U.S. national security laboratories: Lawrence Livermore National Laboratory, Los Alamos National Laboratory, and Sandia National Laboratories.
VIEWS	Visual Interactive Environment for Weapons Simulation: An ASCI project responsible for the development of a software infrastructure which enables the interaction and visualization of ASCI scale datasets. VIEWs software will permit seeing and understanding the results of ASCI codes.
WAN	Wide Area Network: Any network technology that spans large geographic distances. ASCI WANs must be able to span Northern California and New Mexico with high speed links, and possibly other sites with lower speed links. (Contrast with Local Area Network and Metropolitan Area Network.)
White	A 12 Tflop IBM SP at LLNL. See http://www.llnl.gov/asci/news/white_news.html